Rillor

METHODS NOTE

How to specify a training dataset before anyone builds it

How to create a dataset for LLM training starts with a written specification: the use, sources and rights, fields and labels, format, coverage, size, held-out data and acceptance tests.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Published templates, audits and experiments show what a training dataset specification has to fix before collection starts, and what goes wrong when a line is left open.

- Writing data requirements before collection is published practice. Datasheets for Datasets asks creators to read its composition questions "prior to any data collection" [1], and a dataset requirements specification comes first of five documents adapted from software development [2].
- Data problems often surface late. In interviews with 53 practitioners in high-stakes fields, 92% reported at least one data cascade, a chain of compounding downstream effects from data issues, and 45.3% reported two or more in a project [3].
- Licenses and terms need checking at the original source. An audit of 1,858 fine-tuning datasets found over 70% of licenses for popular datasets on two hosting platforms unspecified [9], and restrictions on web sources rose quickly within one year, 2023 to 2024 [10].
- How records and labels are written belongs in the specification. In one dataset where each example pairs two sentences, a classifier that read only the second sentence predicted the label about 67% of the time [13], and ten widely used test sets carry at least 3.3% label errors on average [12].
- Size depends on the job. One model provider's fine-tuning guide reports gains from 50 to 100 examples [15], one study tuned a 65-billion-parameter model on 1,000 curated examples [21], and pretraining estimates that balance model size and data for a fixed compute budget work out to about 20 tokens per parameter (derived) [22].
- Evaluation data is fixed and deduplicated in the specification. In one study, 4.6% of a web corpus's validation set had a near duplicate in its training set [24], and a survey of 17 fields found errors affecting 329 papers, with leakage, such as duplicates across splits, in every one of those fields [25].
- Acceptance criteria need a measure, a threshold and a sample size. A reviewed sample of 100 records with no errors still allows a true error rate up to 3.62% at 95% confidence (two-sided, exact binomial method, derived) [30].

Background

The document this note is about

Hutchinson and colleagues (FAccT 2021) propose five documents for accountable dataset development: a requirements specification, a design document, an implementation diary, a testing report and a maintenance plan, each "directly analogous to documentation types produced by the Software Development Lifecycle" [2]. They note, citing earlier work, that "machine learning is characterized by a lack of both requirements specifications and design specifications" [2].

Document	The question it answers (Table 1 of [2])
Dataset requirements specification	Is data needed, by whom, for what intended uses, with what properties, and do the uses constrain collection or annotation?
Dataset design document	Is a new dataset needed, how are the requirements put into practice, and what tradeoffs and assumptions are made?
Dataset testing report	Does the dataset meet its requirements?
Dataset maintenance plan	How is stale data detected and fixed, and how are errors fixed?

Their Appendix A template for the specification runs from vision, motivation and intended and non-intended uses through data mocks, creation, instance, distributional, processing, performance, maintenance and sharing requirements to caveats, data ethics, a sign-off grid and a changelog [2]. The data mocks ask for "2-3 typical examples" of what the records should look like [2].

Other documentation frameworks

Datasheets for Datasets (Communications of the ACM, December 2021) groups its questions in seven sections: motivation, composition, collection process, preprocessing, cleaning and labeling, uses, distribution and maintenance [1]. Counted for this note, they hold 57 questions (derived), and the authors say not every question applies to every dataset [1]. Creators "should read through these questions prior to any data collection and then provide answers once data collection is complete" [1].

Data Cards (FAccT 2022) consolidate recurring questions into 31 themes whose importance varies by task, and a survey of 191 respondents found it also varies by dataset modality and job function [4]. Data statements for language data (2018) open with a curation rationale, "Which texts were included and what were the goals in selecting texts", then record the language variety with a BCP-47 tag such as en-US [5]. Their authors cite surveys in which 33 to 81% of papers failed to name the language studied [5].

Why the order matters

Sambasivan and colleagues (CHI 2021) interviewed 53 practitioners in high-stakes fields in India, the United States and East and West African countries [3]. They define data cascades as "compounding events causing negative, downstream effects from data issues, that result in technical debt over time" [3]. Of the 53, 92% reported at least one cascade and 45.3% two or more in a project [3]. The cascades were opaque: practitioners judged the whole system through proxy metrics such as accuracy instead of measuring the data [3]. The study's Table 2 lists four triggers, with poor documentation across organizations at 20.8% [3]. The abstract calls cascades "pervasive (92% prevalence), invisible, delayed, but often avoidable" [3].

This note is for teams that commission or build a dataset for training or evaluating an LLM. It stops at the specification and the acceptance test: what is written down before the build, and how the delivered data is checked against it.

Method

What was examined. Documentation templates for datasets, legal and standards texts on training data, audits of licensing and web consent, studies of label artifacts, data age, language coverage, filtering, deduplication and leakage, studies of dataset size, format documentation, and guidance on acceptance sampling. Each was read for what a specification should state and what goes wrong without it.

Sources. Thirty-one sources, each opened on 7 November 2026: 13 peer-reviewed papers, 8 preprints cited in the arXiv versions listed, 6 legal, regulatory, standards and reference documents (Article 10 of the EU AI Act, the European Commission's template for public summaries of training content, NIST AI RMF, a preview of ISO/IEC 5259-2 and two sections of the NIST/SEMATECH e-Handbook of Statistical Methods) and 4 technical documents (one model provider's fine-tuning guide, one open-source training library's documentation, the Croissant specification and the Open Data Contract Standard).

How. Each line of a specification is tied to the sources that show what it prevents. Figures marked "derived" are arithmetic on a source's own numbers, including upper limits on error rates computed with the exact binomial method NIST describes [30]. Laws are restated as they read and for the systems they cover. This note is research and gives no legal advice. No product, vendor or dataset is ranked.

Where sources need care. The Data Provenance Initiative gives the unspecified license share as "70%+" and "72%+" in different places [9]. The data age study writes "2.8 on average" for temporal degradation between fine-tuning and evaluation data, while its Table 2 gives means of 2.36 and 2.84 for its two model sizes [18]. Article 10 is cited from the European Commission's AI Act Service Desk, whose text is based on the consolidated version as at 27 July 2026 and marks the Digital Omnibus changes [6]. Only paragraphs 2 to 4 are cited.

Findings

1. The specification starts from the use

Every template opens with purpose. Datasheets asks "For what purpose was the dataset created? Was there a specific task in mind?" and asks creators to reflect on the tasks for which the dataset should and should not be used [1]. Hutchinson's template separates intended from non-intended uses, and its design document asks which existing datasets are related and why they are unsuitable [2].

ISO/IEC 5259-2 puts the same order in a standard. "A data usage scope describes how and where the data can be used in an analytics or ML task", and a data quality requirement sets "acceptance criteria relative to the data usage scope", which "can be quantitative or qualitative" [8].

For high-risk systems, Article 10(2) of the EU AI Act requires data governance practices "appropriate for the intended purpose", covering among other things design choices, collection processes and the origin of data, preparation such as annotation, labelling and cleaning, and "an assessment of the availability, quantity and suitability of the data sets that are needed" [6]. Article 10(3) requires the data sets to be "relevant, sufficiently representative" and, as far as possible, "free of errors and complete in view of the intended purpose" [6]. Read as a list, these are lines of a specification, binding only on the systems the Act covers [6]. NIST AI RMF MAP 2.3 asks that data collection and selection considerations, "e.g., availability, representativeness, suitability", be identified and documented [7].

One model provider's fine-tuning guide puts evaluation ahead of data. It advises fine-tuning only after evaluations are set up, because a team needs "a reliable way to determine whether your fine-tuned model is performing better than a base model" [15]. It asks for examples that match "Whatever prompts and outputs you expect in your application, as realistic as possible" [15].

2. The source universe and the rights to each source

Hutchinson's creation requirements ask where the data should come from, with "sources and collection methods", and its sharing requirements cover licensing, access, usage and distribution [2].

Licenses need reading at the original source. The Data Provenance Initiative audited 44 fine-tuning collections made of 1,858 datasets [9]. It found over 70% of licenses for popular datasets on two hosting platforms "Unspecified" [9]. Across three platforms, the listed license matched the audited one 43%, 35% and 54% of the time and was too permissive 29%, 27% and 16% of the time [9]. The authors trace part of the problem to contributors who mistook licenses attached to code repositories for licenses attached to the data [9].

Terms change within a year. Consent in Crisis audited 14,000 web domains and found "a rapid crescendo of data restrictions from web sources" in 2023 to 2024, leaving roughly 5% or more of all tokens in the C4 corpus, or "28%+ of the most actively maintained, critical sources in C4", fully restricted [10]. Under terms of

service restrictions, 45% of C4 was restricted, and the audit found inconsistencies between websites' terms and their robots.txt files [10]. A source list therefore carries the date each source's terms were read.

The European Commission's template for public summaries of training content, written for Article 53(1)(d), sorts sources into publicly available datasets, private datasets obtained from third parties, data crawled from online sources, user data, synthetic data and other sources [11]. For crawled data it asks for the purposes, crawler behaviour and the period of collection [11]. A specification can use the same categories whether or not the template applies to the reader.

Synthetic records need their own line. Shumailov and colleagues found that "indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear" [28]. A specification states whether synthetic records are allowed, their share and how they were made.

3. Records, entities, fields and labels

Hutchinson's instance requirements ask for "numbers and types of instances, features, and labels", and its processing requirements ask how the data should be annotated, filtered and validated [2]. Datasheets asks what each instance represents, how many there are of each type and whether a sample is representative of a larger set [1]. It notes that "any dataset containing text that was written by people relates to people" [1]. The record schema names each entity type that must resolve and join across sources, each required field with its type and each label with its definition, and the mock records make those definitions concrete [2].

Labeling instructions shape the data. In two natural language inference datasets, the writing protocol "leaves clues that make it possible to identify the label by looking only at the hypothesis" [13]. A simple text classifier that read only the hypothesis was right for about 67% of SNLI and 53% of MultiNLI [13].

The writers leave a trace too. Geva and colleagues found that models improved when given annotator identifiers and generalized poorly to annotators absent from training [14]. They recommend that "test set annotators should be disjoint from training set annotators" [14].

Label errors reach test sets. Northcutt and colleagues estimate "an average of at least 3.3% errors" across the test sets of ten widely used datasets, with at least 6% in the ImageNet validation set [12]. On data with many label errors, lower-capacity models may be more useful in practice than higher-capacity ones [12]. A label accuracy target with a review sample gives the delivered set a measured error rate.

4. Format follows the training method

The format line states the record layout a pipeline reads. One model provider's guide asks for JSON Lines, "with one complete JSON structure on every line of the training data file", in its chat format, in a file that "must have at least 10 lines" [15].

One open-source training library separates format from type. Format is "either standard or conversational", type is tied to the task, "such as prompt-only or preference", and each type is "characterized by its columns" [16]. Its types include language modeling, prompt-only, prompt-completion, preference records with "prompt", "chosen" and "rejected", and unpaired preference records with a boolean label [16].

Croissant, an MLCommons specification of 1 March 2024, describes the delivered files in JSON-LD [17]. A data record is "A granular part of a dataset, such as an image, text file, or a row in a table", and a record set is "A set of homogeneous data records" [17]. Its examples give each file a content URL, an encoding format and a sha256 checksum, and each record set typed fields [17].

Format line	What it states	Example from the sources
File format	Encoding and record boundaries	JSON Lines, one JSON structure per line [15]
Record type	The training task the records serve	Prompt-completion or preference [16]
Columns	Field names for each record type	"prompt", "chosen", "rejected" [16]
Description	File list, checksums and fields	Croissant in JSON-LD with sha256 [17]

5. Coverage: time range, languages and regions

The time range follows the period the model serves. Longpre and colleagues pretrained 28 models of 1.5 billion parameters on data curated at different times and found that "A temporal shift between evaluation data and pretraining data leads to performance degradation, which is not overcome by finetuning" [18]. Their Table 2 puts the average loss from one year of misalignment at 2.36 and 2.84 points for two model sizes when the gap is between fine-tuning and evaluation data, and at 0.08 and 0.41 points when it is between pretraining and evaluation data [18]. They also found heterogeneous sources, such as books and web text, broadly beneficial [18].

Language labels need checking on a read sample. Kreutzer and colleagues audited 205 language-specific corpora from five public datasets: "At least 15 corpora have no usable text", a significant fraction held less than 50% acceptable sentences, and many were mislabeled [19]. The issues were "easy to detect even for non-proficient speakers" [19]. Data statements supply the field: a BCP-47 tag plus a prose description of the variety [5].

A corpus's description can mislead. Dodge and colleagues found that C4, built by filtering one Common Crawl snapshot, holds text "from unexpected sources like patents and US military websites" and machine-generated text, and is "by no means representative of English-speaking world" [20].

For regions, Datasheets asks whether a sample is representative, with geographic coverage as its example [1]. For high-risk systems, Article 10(4) requires data sets to take into account, "to the extent required by the intended purpose", the "geographical, contextual, behavioural or functional setting" of use [6].

6. Cleaning and filtering rules are hypotheses to test

Each filter changes the coverage the specification promises. In a sample of 100,000 documents removed from C4 by its blocklist, only 16 clusters, 31% of the removed documents, were largely sexual, and others concerned science, medicine and health, and legal and political documents [20]. The filter removed African American English and Hispanic-aligned English at 42% and 32%, against 6.2% and 7.2% for White-aligned and other English [20].

Deduplication needs the same testing. The FineWeb authors compared design choices with identical ablation models of 1.71 billion parameters, trained on about 28 billion tokens for filtering ablations and on 350 billion tokens for some deduplication ablations [27]. Global MinHash deduplication across 96 snapshots removed up to 90% of the oldest snapshots' data, yet the models showed little improvement over non-deduplicated data [27]. For the 2013-48 crawl, the 10% kept "was actually of worse quality than the 90% of data that was removed" [27]. Deduplicating each snapshot on its own left 20 trillion tokens and matched the performance of another public web dataset [27].

Other sources point the same way. The data age study found no "one-size-fits-all solution to filtering training data" [18]. Lee and colleagues found near duplicates in 3.04% of C4 training examples and 13.63% of RealNews, and their deduplicated models emitted memorized text ten times less often [24]. The LIMA study wrote its filter as a rule, dropping answers under 1,200 or over 4,096 characters or written in the first person [21].

The specification states each filter and deduplication rule, and asks for a reviewed sample of what each removed.

7. Size tied to the job

For fine-tuning, one model provider's guide sets a minimum of 10 examples, reports improvements from 50 to 100 and says the right number "varies greatly and depends on the use case" [15]. It suggests starting with 50 well-crafted demonstrations, and if those have no impact, rethinking the task or prompt before adding data [15].

Zhou and colleagues fine-tuned a 65-billion-parameter model "on only 1,000 carefully curated prompts and responses", roughly 750,000 tokens, or about 750 tokens per example (derived) [21]. In ablations on a 7-billion-parameter model, diverse data beat homogeneous data at 2,000 examples each, filtered data beat unfiltered data by 0.5 points on a 1 to 6 scale graded by a model, and "doubling the training set does not improve response quality" [21]. They conclude that results depend on "prompt diversity while maintaining high quality responses" more than quantity alone [21].

Pretraining sizes follow the model and the compute budget. Hoffmann and colleagues trained over 400 models and found that "for every doubling of model size the number of training tokens should also be doubled" [22]. Their Table 3 estimates 20.2 billion tokens for 1 billion parameters and 1.5 trillion for 67 billion, a ratio of 20 to 22.4 tokens per parameter across its rows (derived) [22]. When unique data runs short, Muennighoff and colleagues found that "training with up to 4 epochs of repeated data yields negligible changes to loss compared to having unique data", with returns decaying beyond that [23].

A size line therefore states a range, unique records and total tokens apart, and a rule for growth: evaluate at a starting size, then add data where the evaluation shows a gap. The European Commission's template reports text size after pre-processing in broad ranges: under 1 billion tokens, 1 billion to 10 trillion, and over 10 trillion [11].

8. Held-out evaluation data, duplicates and leakage

Datasheets asks whether there are recommended splits for training, validation and testing [1]. The specification fixes them before collection, because overlap is common. Lee and colleagues report "train-test overlap, which affects over 4% of the validation set of standard datasets" [24]. In their data, 4.6% of the C4 validation set and 14.4% of the RealNews validation set had an approximate duplicate in training, which "could cause evaluation metrics to be unfairly inflated" [24]. When text appeared in both, they kept the evaluation copy and removed the training one [24].

Benchmark text sits inside web corpora. In C4's English set, 4.6% and 5.7% of the examples in two benchmark test sets appear verbatim, and the share of GLUE test inputs found there varies from under 2% to over 50% [20].

Leakage takes many forms. Kapoor and Narayanan found errors in 17 fields, collectively affecting 329 papers, with leakage in every field, and sorted leakage into eight types that include duplicates in datasets, preprocessing on training and test sets together and temporal leakage [25]. They propose model info sheets that explain how the training and test sets are split during all steps [25].

The split also needs enough records. Card and colleagues found that "typical test sets of 2000 sentences have approximately 75% power to detect differences of 1 BLEU point" [26]. The specification states the smallest difference the evaluation needs to detect and sizes the split for it. It also keeps the writers of test records apart from those of training records [14] and matches the evaluation period to the period of use [18].

9. Acceptance criteria and sampling

Hutchinson and colleagues describe requirements testing, also called acceptance testing, which "checks a dataset against the stated requirements for its use" [2]. Ideal data testing "returns boolean pass/fail outcomes", and difficulty in reaching one "may point to a lack of clarity in the dataset requirements" [2]. Requirements need to be traceable, and in many cases measurable [2].

ISO/IEC 5259-2 lists 23 data quality characteristics in four groups (counted from its contents page), among them accuracy, completeness, consistency, currentness, traceability, balance, diversity, representativeness and timeliness [8].

Acceptance sampling suits datasets too large to review in full. NIST describes it as drawing a random sample from a lot and deciding from it whether to accept or reject the lot, "the middle of the road" between no inspection and 100% inspection [29]. Its purpose "is to decide whether or not the lot is likely to be acceptable, not to estimate the quality of the lot" [29].

A clean sample still leaves an upper limit on the true error rate. With the exact binomial method NIST describes for small samples and few failures [30], the two-sided 95% upper limits for zero errors are:

Records reviewed, zero errors	Upper 95% limit on the error rate (derived)
50	7.11%
100	3.62%
200	1.83%
500	0.74%
1,000	0.37%

Source: computed with the method in [30].

A criterion reads as a measure, a threshold, a sample size and the interval the sample supports. Measurable lines include the completeness of required fields, label accuracy against a second review, the near-duplicate rate against the held-out split and language identification on a read sample [2, 12, 19, 24].

Data contracts write such rules in a checkable form. The Open Data Contract Standard, whose sections include a schema, data quality rules and a service-level agreement, lets quality rules be plain text, library metrics such as row counts and null values, SQL queries or custom checks [31]. Library metrics "come with an operator to compare if the metric is valid and in the expected range" [31]. A dataset specification can borrow that form.

10. Refresh, versions and the documents delivered

Datasheets asks whether the dataset is updated to correct labeling errors, add instances or delete them, how often and by whom, and whether older versions stay supported [1]. Hutchinson's maintenance requirements ask how often the data is updated and how long it is retained, and its maintenance plan asks

how stale data and errors are detected and fixed [2].

Findings 2 and 5 set the cadence. Web terms changed within one year [10], and a time gap between training and evaluation data lowered performance [18]. A refresh line names what triggers a new version: changed terms, withdrawn material, a new period of data or a correction. Each version ships with its documents: a changelog [2], a machine-readable description of files and fields [17] and a summary for stakeholders "across a dataset's lifecycle" [4].

Implications

A specification checklist. Each line states what is fixed before the build and how the delivered data is tested against it.

Line	What it states	How acceptance tests it	Basis
Use	Task, intended and non-intended uses, the evaluation that measures improvement	Evaluation set ready before training	[1, 2, 8, 15]
Source universe	Sources in and out, category, license read at the source, date read, synthetic share	Each source has a recorded license and date	[9, 10, 11, 28]
Entities and fields	Entity types that resolve and join, required fields, two or three mock records	Required-field completeness on a reviewed sample	[1, 2]
Labels	Definitions, writing instructions, separate writers for test and training records	Label accuracy against a second review	[12, 13, 14]
Format	Record type, columns, file format, file description	Files parse and checksums match	[15, 16, 17]
Coverage	Time range, languages with tags, regions	Read sample for language and date	[5, 18, 19]
Cleaning	Each filter and deduplication rule	Reviewed sample of removed records	[20, 24, 27]
Size	A range, unique records and total tokens apart, a rule for growth	Counts at delivery	[15, 21, 22, 23]
Held-out data	Splits fixed before collection, deduplicated against training, sized for power	Near-duplicate check across splits	[24, 25, 26]
Acceptance	Measure, threshold, sample size and interval for each criterion	Accept or reject on a random sample	[2, 29, 30]
Refresh and versions	Cadence, triggers, version history, documents per version	Changelog for each version	[1, 2, 17]

For teams commissioning a dataset. Write the use and the evaluation before the source list, since every later line is measured against them [2, 8, 15]. Ask for each source's license as read at the original source, with the date [9, 10]. Set size as a range with a rule for growth, and weigh diversity and quality alongside volume [15, 21]. Read a clean acceptance sample as an upper limit on the error rate [29, 30].

For teams building it. Keep the mock records, filters and split rules in the same versioned document as the schema. Review a sample of what each filter removes, since filters can remove some dialects at several times the rate of others [20] and can keep worse data than they drop [27]. Deduplicate the evaluation split against training and keep the evaluation copy [24]. Record each version's files, checksums and fields in a machine-readable description [17].

For reviewers. A requirement that cannot be tested pass or fail points to an unclear specification [2].

Limits of this analysis

- This note covers what a specification states and how acceptance is tested. It does not cover how to collect, label or clean data, and it ranks no product, vendor or dataset.
- The data cascades study interviewed 53 practitioners in high-stakes fields and was not specific to LLM training [3].
- Many findings come from English-heavy research datasets. They show what each line of a specification guards against, never rates to expect in a new dataset.
- The LIMA results come from one model family and helpfulness tasks, and its data ablations were graded by a model [21]. No source reviewed gives a general rule for how many examples a task needs.
- The fine-tuning guide cited states that its platform is no longer open to new users and that eligible existing customers can create training jobs until 6 January 2027 [15]. Its guidance is cited as published, not as a platform to choose.
- Article 10 of the EU AI Act covers high-risk systems only [6], and the Commission's template covers providers of general-purpose models [11]. Both are restated as written, never as legal advice or a statement of anyone's compliance.
- Only the preview of ISO/IEC 5259-2 was read [8].
- No source reviewed gives a cost multiplier for fixing data problems late, a time or cost for building a dataset, or a target for label agreement.
- The derived intervals treat the reviewed records as independent random draws [30].

Sources

1. Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione and others, Communications of the ACM, arXiv 1803.09010, version 8, 1 December 2021, viewed 7 November 2026. https://arxiv.org/abs/1803.09010

2. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. Ben Hutchinson, Andrew Smart, Alex Hanna and others, FAccT 2021, arXiv 2010.13561, version 2, 30 January 2021, viewed 7 November 2026. https://arxiv.org/abs/2010.13561

3. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Nithya Sambasivan, Shivani Kapania, Hannah Highfill and others, CHI 2021, Google Research, May 2021, viewed 7 November 2026. https://research.google/pubs/everyone-wants-to-do-the-model-work-not-the-data-work-data-cascades-in-high-stakes-ai/

4. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. Mahima Pushkarna, Andrew Zaldivar and Oddur Kjartansson, FAccT 2022, arXiv 2204.01075, 3 April 2022, viewed 7 November 2026. https://arxiv.org/abs/2204.01075

5. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Emily M. Bender and Batya Friedman, Transactions of the Association for Computational Linguistics, volume 6, 2018, viewed 7 November 2026. https://aclanthology.org/Q18-1041/

6. Regulation (EU) 2024/1689 (AI Act), Article 10: Data and data governance, consolidated text as at 27 July 2026. European Commission, AI Act Service Desk, viewed 7 November 2026. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-10

7. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. National Institute of Standards and Technology, January 2023, viewed 7 November 2026. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

8. ISO/IEC 5259-2:2024, Artificial intelligence: Data quality for analytics and machine learning (ML), Part 2: Data quality measures, first edition. ISO/IEC, November 2024, preview sample via iTeh Standards, viewed 7 November 2026. https://cdn.standards.iteh.ai/samples/81860/6ac019c359644309afa716572bb73edc/ISO-IEC-5259-2-2024.pdf

9. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Shayne Longpre, Robert Mahari, Anthony Chen and others, arXiv 2310.16787, version 3, 4 November 2023, viewed 7 November 2026. https://arxiv.org/abs/2310.16787

10. Consent in Crisis: The Rapid Decline of the AI Data Commons. Shayne Longpre, Robert Mahari, Ariel Lee and others, arXiv 2407.14933, version 2, 24 July 2024, viewed 7 November 2026. https://arxiv.org/abs/2407.14933

11. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, AI Office, 24 July 2025 (Communication C(2025) 8311 final, 5 December 2025), viewed 7 November 2026. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models

12. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. Curtis G. Northcutt, Anish Athalye and Jonas Mueller, NeurIPS 2021 Datasets and Benchmarks, arXiv 2103.14749, version 4, 7 November 2021, viewed 7

November 2026. https://arxiv.org/abs/2103.14749

13. Annotation Artifacts in Natural Language Inference Data. Suchin Gururangan, Swabha Swayamdipta, Omer Levy and others, NAACL 2018, arXiv 1803.02324, version 2, 16 April 2018, viewed 7 November 2026. https://arxiv.org/abs/1803.02324

14. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. Mor Geva, Yoav Goldberg and Jonathan Berant, EMNLP-IJCNLP 2019, arXiv 1908.07898, version 2, 28 August 2019, viewed 7 November 2026. https://arxiv.org/abs/1908.07898

15. Supervised fine-tuning. OpenAI, API documentation, viewed 7 November 2026. https://developers.openai.com/api/docs/guides/supervised-fine-tuning

16. Dataset formats and types. Hugging Face, TRL documentation, viewed 7 November 2026. https://huggingface.co/docs/trl/dataset_formats

17. Croissant Format Specification, version 1.0. Omar Benjelloun, Elena Simperl, Pierre Marcenac and others, MLCommons, 1 March 2024, viewed 7 November 2026. https://docs.mlcommons.org/croissant/docs/croissant-spec.html

18. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. Shayne Longpre, Gregory Yauney, Emily Reif and others, arXiv 2305.13169, version 2, 13 November 2023, viewed 7 November 2026. https://arxiv.org/abs/2305.13169

19. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Julia Kreutzer, Isaac Caswell, Lisa Wang and others, Transactions of the Association for Computational Linguistics 10, 2022, arXiv 2103.12028, version 4, 21 February 2022, viewed 7 November 2026. https://arxiv.org/abs/2103.12028

20. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. Jesse Dodge, Maarten Sap, Ana Marasović and others, EMNLP 2021, arXiv 2104.08758, version 2, 30 September 2021, viewed 7 November 2026. https://arxiv.org/abs/2104.08758

21. LIMA: Less Is More for Alignment. Chunting Zhou, Pengfei Liu, Puxin Xu and others, arXiv 2305.11206, 18 May 2023, viewed 7 November 2026. https://arxiv.org/abs/2305.11206

22. Training Compute-Optimal Large Language Models. Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch and others, arXiv 2203.15556, 29 March 2022, viewed 7 November 2026. https://arxiv.org/abs/2203.15556

23. Scaling Data-Constrained Language Models. Niklas Muennighoff, Alexander M. Rush, Boaz Barak and others, arXiv 2305.16264, version 5, 28 June 2025, viewed 7 November 2026. https://arxiv.org/abs/2305.16264

24. Deduplicating Training Data Makes Language Models Better. Katherine Lee, Daphne Ippolito, Andrew Nystrom and others, ACL 2022, arXiv 2107.06499, version 2, 24 March 2022, viewed 7 November 2026. https://arxiv.org/abs/2107.06499

25. Leakage and the Reproducibility Crisis in ML-based Science. Sayash Kapoor and Arvind Narayanan, arXiv 2207.07048, 14 July 2022, viewed 7 November 2026. https://arxiv.org/abs/2207.07048

26. With Little Power Comes Great Responsibility. Dallas Card, Peter Henderson, Urvashi Khandelwal and others, EMNLP 2020, arXiv 2010.06595, 13 October 2020, viewed 7 November 2026. https://arxiv.org/abs/2010.06595

27. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal and others, arXiv 2406.17557, version 2, 31 October 2024, viewed 7 November 2026. https://arxiv.org/abs/2406.17557

28. AI models collapse when trained on recursively generated data. Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao and others, Nature 631, 755-759, 24 July 2024, viewed 7 November 2026. https://www.nature.com/articles/s41586-024-07566-y

29. NIST/SEMATECH e-Handbook of Statistical Methods, section 6.2.1: What is Acceptance Sampling? National Institute of Standards and Technology, viewed 7 November 2026. https://www.itl.nist.gov/div898/handbook/pmc/section2/pmc21.htm

30. NIST/SEMATECH e-Handbook of Statistical Methods, section 7.2.4.1: Confidence intervals. National Institute of Standards and Technology, viewed 7 November 2026. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

31. Open Data Contract Standard, Data Quality. Bitol, version 3, viewed 7 November 2026. https://bitol-io.github.io/open-data-contract-standard/latest/data-quality/

About this report

Published by Rillor on 8 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. How to specify a training dataset before anyone builds it. 8 November 2026. https://rillor.com/insights/specifying-a-training-dataset

Work with Rillor

Send the question the data should answer, the sources you know of, the coverage you need and how often it should update. Contact us at rillor.com/contact?type=custom-data or contact@rillor.com. More on data services: rillor.com/data